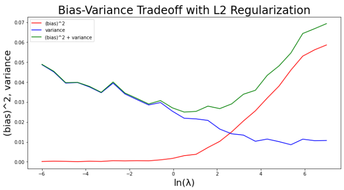
Introduction to Machine Learning

Regularization
Perspectives on Ridge Regression (Deep-Dive)



Learning goals

- Know interpretation of $L2$ regularization as row-augmentation
- Know interpretation of $L2$ regularization as minimizing risk under feature noise
- Bias-variance tradeoff for ridge regression

PERSPECTIVES ON *L*2 REGULARIZATION

We already saw two interpretations of *L*2 regularization.

- We know that it is equivalent to a constrained optimization problem:

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} \quad = \quad \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \lambda \|\boldsymbol{\theta}\|_2^2 = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y}$$

  For some t depending on λ this is equivalent to:

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} \quad = \quad \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 \text{ s.t. } \|\boldsymbol{\theta}\|_2^2 \leq t$$

- Bayesian interpretation of ridge regression: For additive Gaussian errors $\mathcal{N}(0, \sigma^2)$ and i.i.d. normal priors $\theta_j \sim \mathcal{N}(0, \tau^2)$, the resulting MAP estimate is $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ with $\lambda = \frac{\sigma^2}{\tau^2}$:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\theta} \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})] = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \frac{\sigma^2}{\tau^2} \|\boldsymbol{\theta}\|_2^2$$

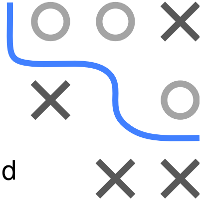
$L2$ AND ROW-AUGMENTATION

We can also recover the ridge estimator by performing least-squares on a **row-augmented** data set: Let $\tilde{\mathbf{X}} := \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\boldsymbol{I}_p \end{pmatrix}$ and $\tilde{\mathbf{y}} := \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_p \end{pmatrix}$. Using the augmented data, the unregularized least-squares solution $\tilde{\boldsymbol{\theta}}$ can be written as

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n+p} \left(y^{\tilde{(i)}} - \boldsymbol{\theta}^T \mathbf{x}^{\tilde{(i)}} \right)^2 \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \sum_{j=1}^{p} \left(0 - \sqrt{\lambda}\theta_j \right)^2 \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \lambda \|\boldsymbol{\theta}\|_2^2
\end{aligned}
$$

$\Longrightarrow \hat{\boldsymbol{\theta}}_{\text{ridge}}$ is the least-squares solution $\tilde{\boldsymbol{\theta}}$ but using $\tilde{\mathbf{X}}, \tilde{\mathbf{y}}$ instead of $\mathbf{X}, \mathbf{y}$!
This is sometimes useful as we can cast the regularized problem as an unregularized one using augmented data

$L2$ AND NOISY FEATURES



Now consider perturbed features $\mathbf{x}^{\tilde{(i)}} := \mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}$ where $\boldsymbol{\delta}^{(i)} \overset{iid}{\sim} (\mathbf{0}, \lambda \boldsymbol{I}_p)$. Note that no parametric family is assumed. We want to minimize the expected squared error taken w.r.t. the perturbations $\boldsymbol{\delta}$:

$$\mathcal{P}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\delta}}\Big[\frac{1}{n}\sum_{i=1}^{n}((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{\tilde{(i)}})^2)\Big] = \mathbb{E}_{\boldsymbol{\delta}}\Big[\frac{1}{n}\sum_{i=1}^{n}((y^{(i)} - \boldsymbol{\theta}^\top (\mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}))^2)\Big] \Big| \text{expand}$$

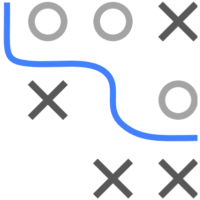
$$\mathcal{P}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\delta}}\Big[\frac{1}{n}\sum_{i=1}^{n}((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 - 2\boldsymbol{\theta}^\top \boldsymbol{\delta}^{(i)}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + \boldsymbol{\theta}^\top \boldsymbol{\delta}^{(i)} \boldsymbol{\delta}^{(i)\top} \boldsymbol{\theta})\Big]$$

By linearity of expectation, $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)}] = \mathbf{0}_p$ and $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)}\boldsymbol{\delta}^{(i)\top}] = \lambda \boldsymbol{I}_p$, this is

$$\begin{aligned}\mathcal{P}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 - 2\boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)}](y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + \boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)}\boldsymbol{\delta}^{(i)\top}]\boldsymbol{\theta}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda \|\boldsymbol{\theta}\|_2^2\end{aligned}$$

$\Longrightarrow$ Ridge regression on unperturbed features $\mathbf{x}^{(i)}$ turns out to be minimizing squared loss averaged over feature noise distribution!

BIAS-VARIANCE DECOMPOSITION FOR RIDGE



For a linear model $\mathbf{y} = \mathbf{X}\theta + \varepsilon$ with fixed design $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\varepsilon \sim (\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, bias of ridge estimator $\hat{\theta}_{\text{ridge}}$ is given by

$$
\begin{aligned}
\text{Bias}(\hat{\theta}_{\text{ridge}}) := \mathbb{E}[\hat{\theta}_{\text{ridge}} - \theta] &= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}] - \theta \\
&= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\theta + \varepsilon)] - \theta \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\theta + (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \underbrace{\mathbb{E}[\varepsilon]}_{=0} - \theta \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\theta - \theta \\
&= \left[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \right] \boldsymbol{X}^\top \boldsymbol{X}\theta
\end{aligned}
$$

- Last expression shows bias of ridge estimator only vanishes for $\lambda = 0$, which is simply (unbiased) OLS solution
- It follows $\|\text{Bias}(\hat{\theta}_{\text{ridge}})\|_2^2 > 0$ for all $\lambda > 0$

BIAS-VARIANCE DECOMPOSITION FOR RIDGE

For the variance of $\hat{\theta}_{\text{ridge}}$, we have

$$
\begin{aligned}
\text{Var}(\hat{\theta}_{\text{ridge}}) &= \text{Var}\left((\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}\right) \qquad \Big| \text{ apply } \text{Var}_u(\mathbf{A}\mathbf{u}) = \mathbf{A}\text{Var}(\mathbf{u})\mathbf{A}^\top \\
&= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \text{Var}(\mathbf{y}) \left((\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top\right)^\top \\
&= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \text{Var}(\varepsilon) \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \\
&= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \sigma^2 \mathbf{I}_n \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \\
&= \sigma^2 (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1}
\end{aligned}
$$

- $\text{Var}(\hat{\theta}_{\text{ridge}})$ is strictly smaller than $\text{Var}(\hat{\theta}_{\text{OLS}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ for any $\lambda > 0$, meaning matrix of their difference $\text{Var}(\hat{\theta}_{\text{OLS}}) - \text{Var}(\hat{\theta}_{\text{ridge}})$ is positive definite (bit tedious derivation)
- This further means $\text{trace}\left(\text{Var}(\hat{\theta}_{\text{OLS}}) - \text{Var}(\hat{\theta}_{\text{ridge}})\right) > 0 \; \forall \lambda > 0$

BIAS-VARIANCE DECOMPOSITION FOR RIDGE

Having obtained the bias and variance of the ridge estimator, we can decompose its mean squared error as follows:

$$\text{MSE}(\hat{\theta}_{\text{ridge}}) = \|\text{Bias}(\hat{\theta}_{\text{ridge}})\|_2^2 + \text{trace}\big(\text{Var}(\hat{\theta}_{\text{ridge}})\big)$$

Comparing MSEs of $\hat{\theta}_{\text{ridge}}$ and $\hat{\theta}_{\text{OLS}}$ and using $\text{Bias}(\hat{\theta}_{\text{OLS}}) = 0$ we find

$$\text{MSE}(\hat{\theta}_{\text{OLS}}) - \text{MSE}(\hat{\theta}_{\text{ridge}}) = \underbrace{\text{trace}\big(\text{Var}(\hat{\theta}_{\text{OLS}}) - \text{Var}(\hat{\theta}_{\text{ridge}})\big)}_{>0} - \underbrace{\|\text{Bias}(\hat{\theta}_{\text{ridge}})\|_2^2}_{>0}$$

Since both terms are positive, sign of their diff is *a priori* undetermined. ▸ Theobald 1974 and ▸ Farebrother 1976 prove there always exists some $\lambda^* > 0$ so that

$$\text{MSE}(\hat{\theta}_{\text{OLS}}) - \text{MSE}(\hat{\theta}_{\text{ridge}}) > 0$$

Important theoretical result: While Gauss-Markov guarantuees $\hat{\theta}_{\text{OLS}}$ is best linear unbiased estimator (BLUE), there are biased estimators with lower MSE.

BIAS-VARIANCE IN PREDICTIONS FOR RIDGE

In supervised learning, our goal is typically not to learn an unknown parameter $\boldsymbol{\theta}$, but to learn a function $f(\mathbf{x})$ that can predict y given $\mathbf{x}$.

The bias and variance of predictions $\hat{f} := \hat{f}(\mathbf{x}) = \hat{\boldsymbol{\theta}}_{\text{ridge}}^\top \mathbf{x}$ is obtained as:

$$
\begin{aligned}
\text{Bias}(\hat{f}) &= \mathbb{E}[\hat{f} - f] = \mathbb{E}[\hat{\boldsymbol{\theta}}_{\text{ridge}}^\top \mathbf{x} - \boldsymbol{\theta}^\top \mathbf{x}] = \mathbb{E}[\hat{\boldsymbol{\theta}}_{\text{ridge}} - \boldsymbol{\theta}]^\top \mathbf{x} \\
&= \text{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}})^\top \mathbf{x} \\
\text{Var}(\hat{f}) &= \text{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}}^\top \mathbf{x}) = \mathbf{x}^\top \text{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) \mathbf{x}
\end{aligned}
$$

The MSE of $\hat{f}$ given a fresh sample $(y, \mathbf{x})$ can now be decomposed as

$$
\text{MSE}(\hat{f}) = \mathbb{E}[(y - \hat{f}(\mathbf{x}))^2] = \text{Bias}^2(\hat{f}) + \text{Var}(\hat{f}) + \sigma^2
$$

This decomposition is similar to the statistical inference setting before, however, the irreducible error σ^2 only appears for predictions as an artifact of the noise in the test sample.